Przetwarzanie języka naturalnego i Text Mining

Dr Bogna Zacny
Katedra Inżynierii Wiedzy
Uniwersytet Ekonomiczny w Katowicach

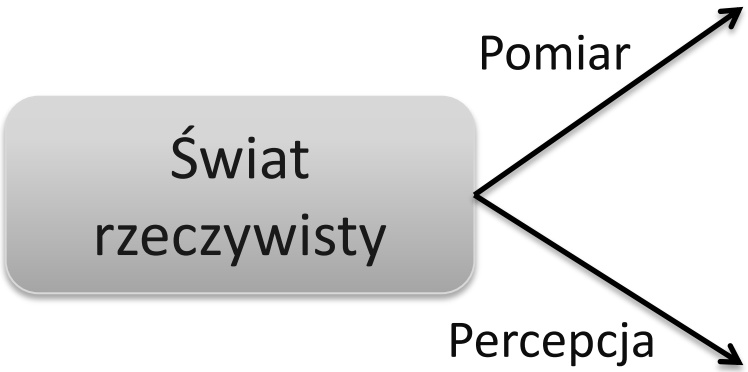
Odkrywanie wiedzy z tekstu

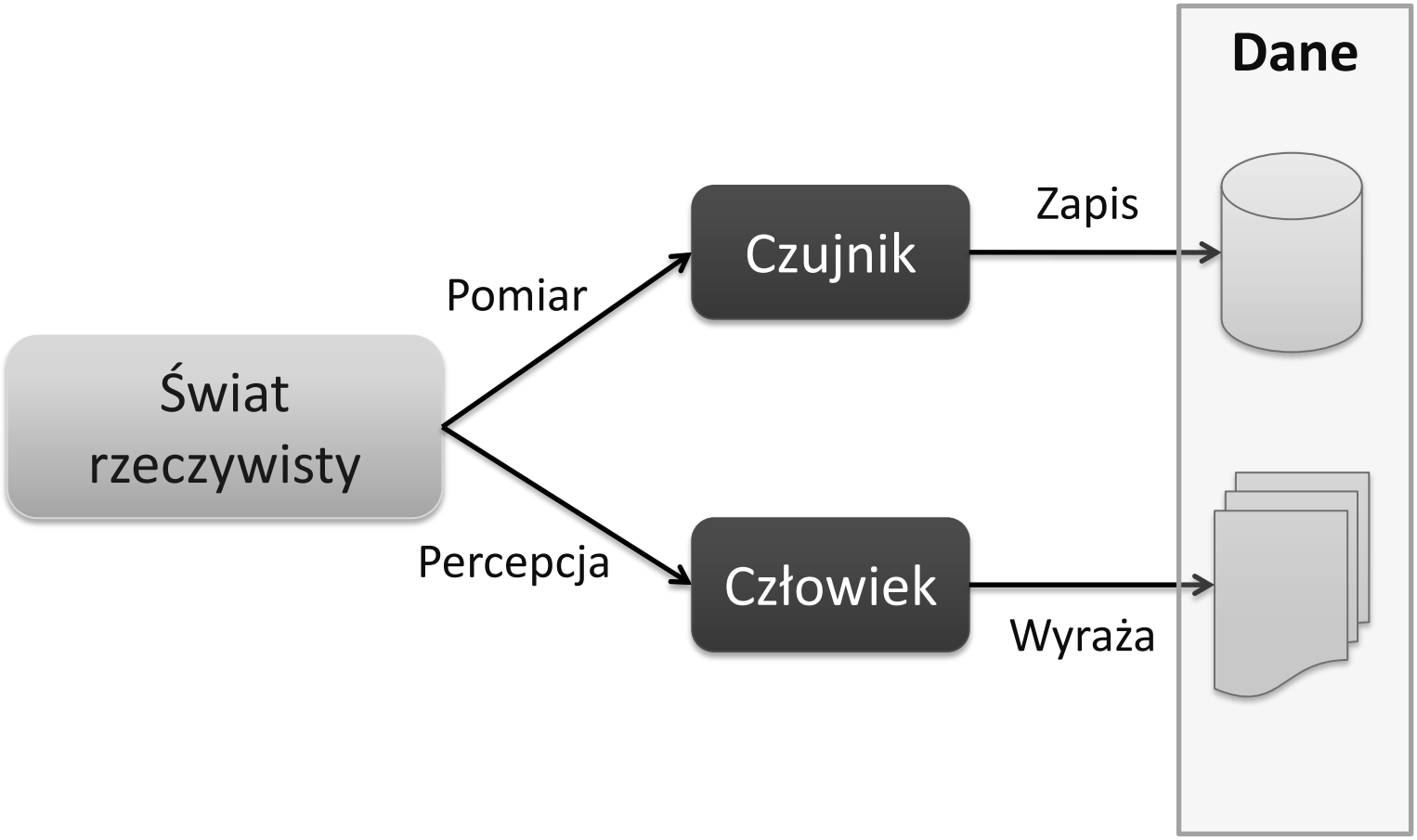
to proces mający na celu wydobycie z zasobów tekstowych nieznanych wcześniej informacji

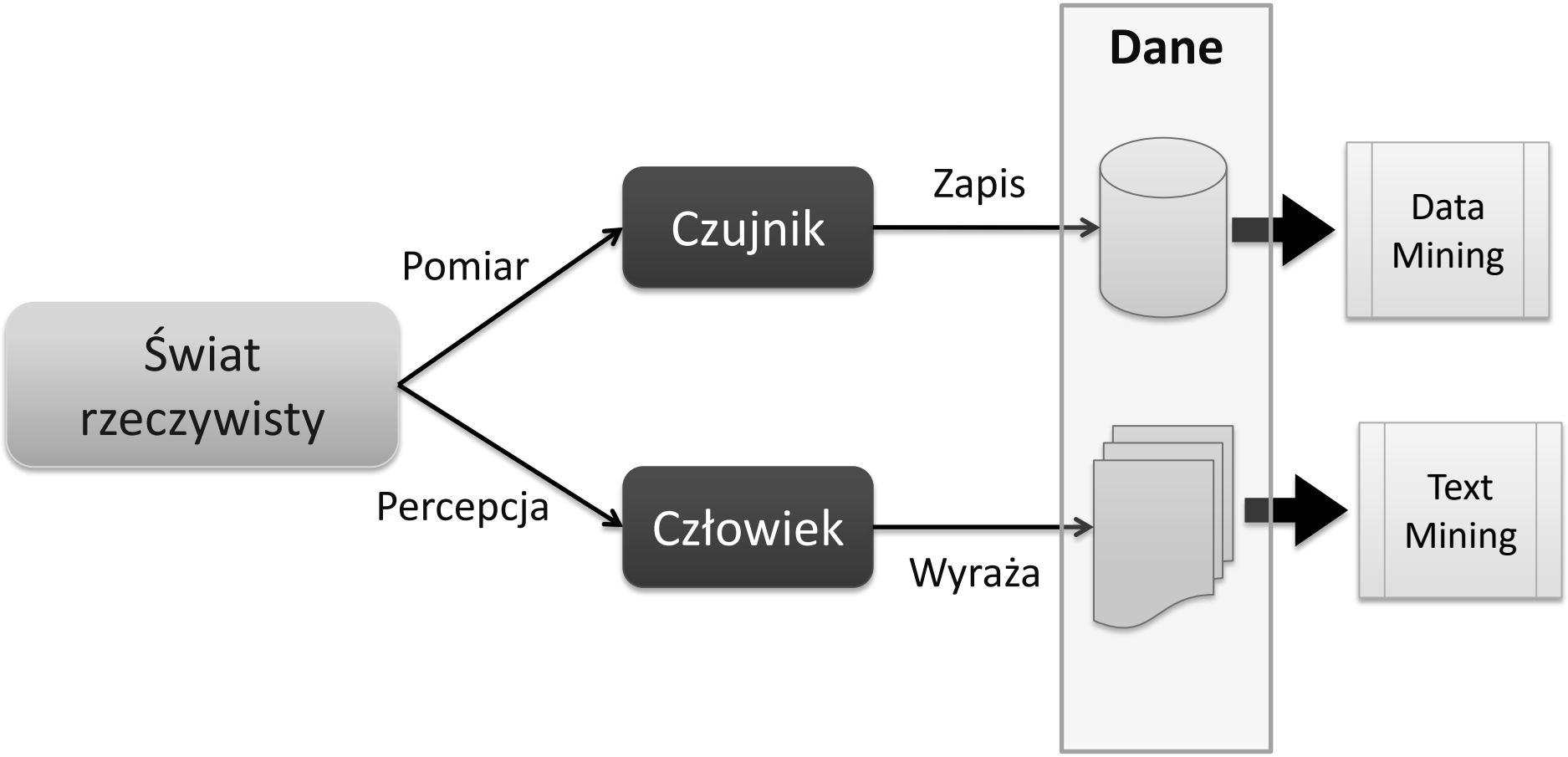
Półautomatyczny proces wydobywania wiedzy z nieustrukturyzowanych źródeł danych

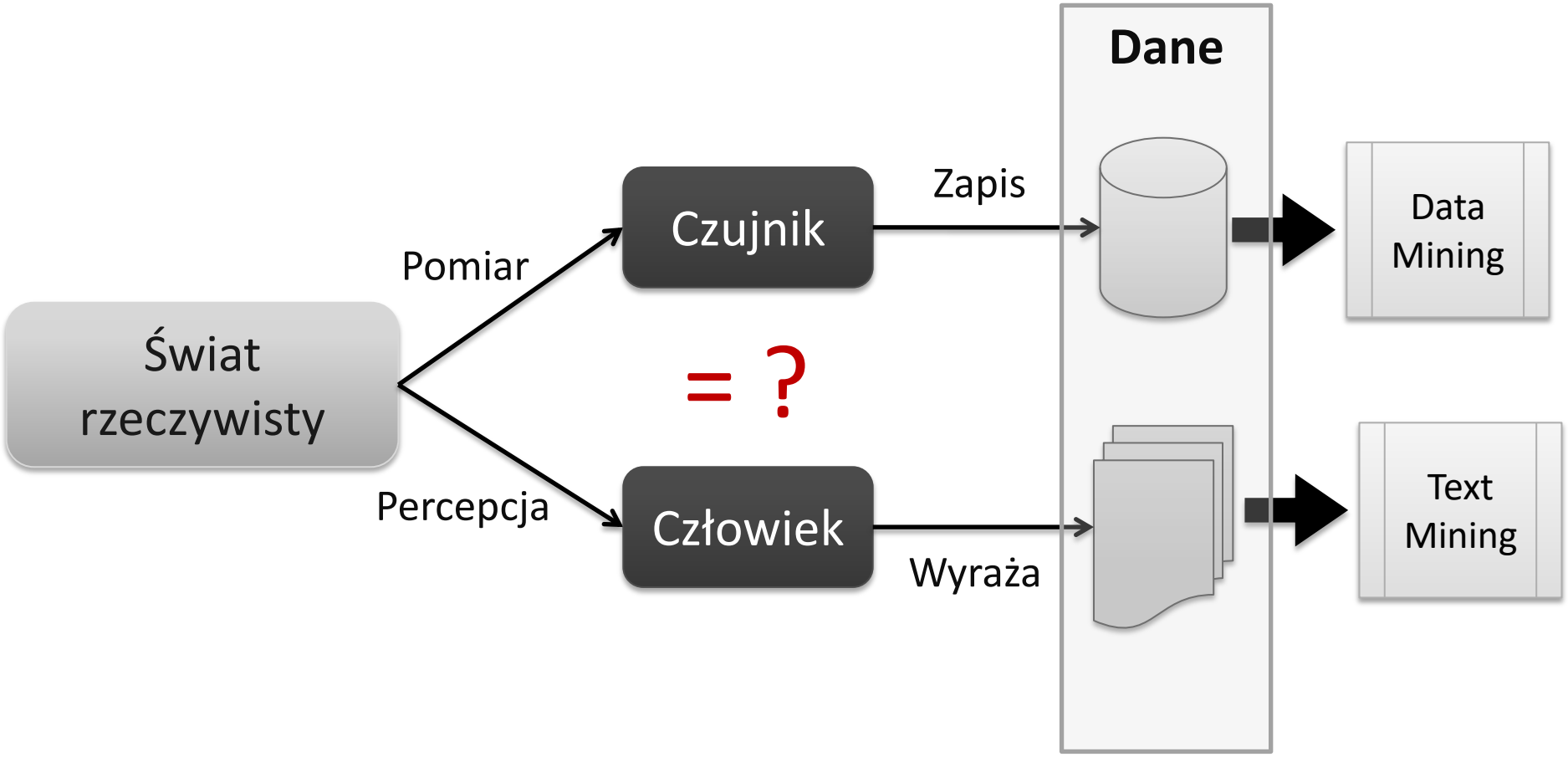
TEKST

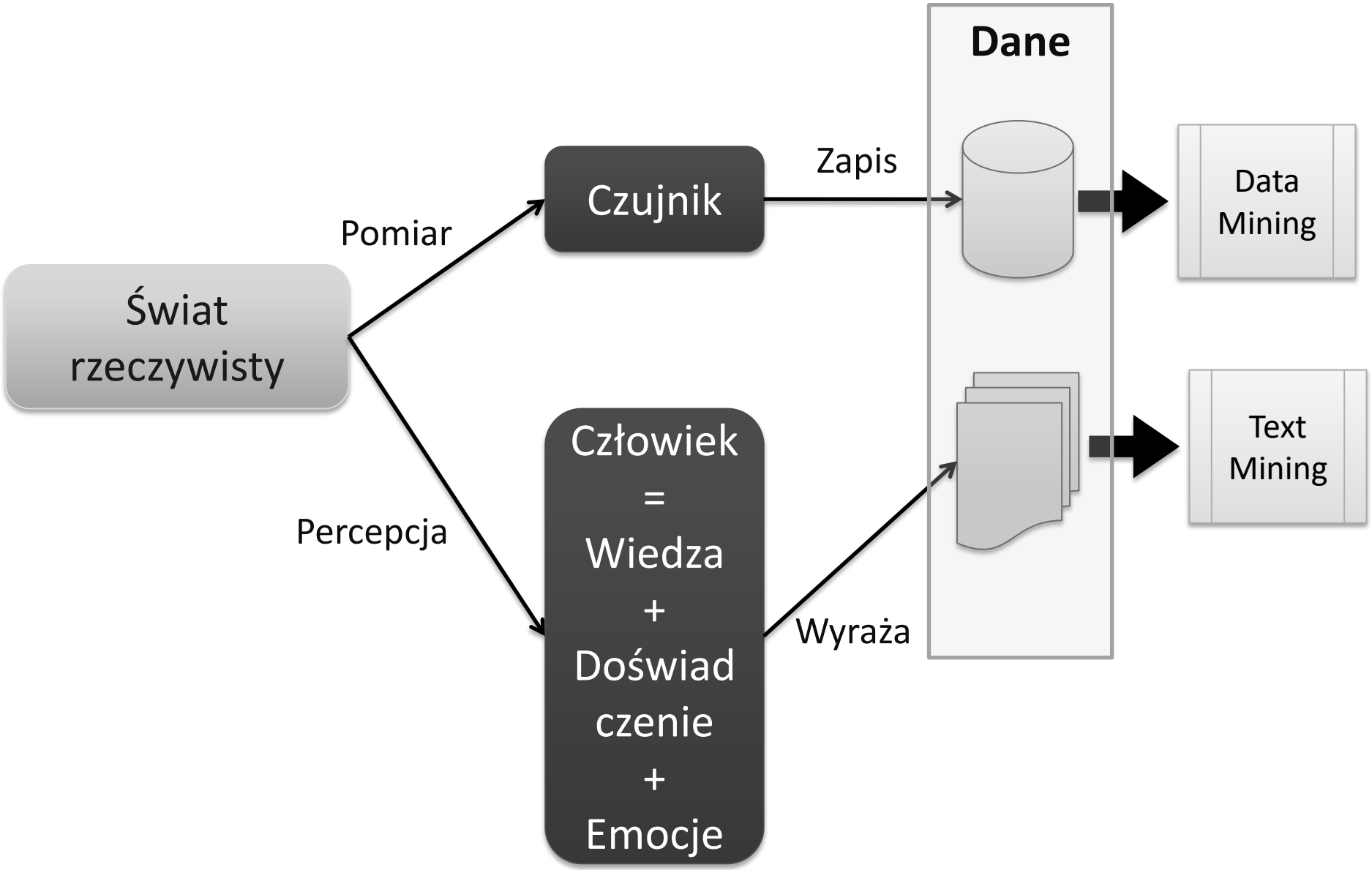
Świat rzeczywisty

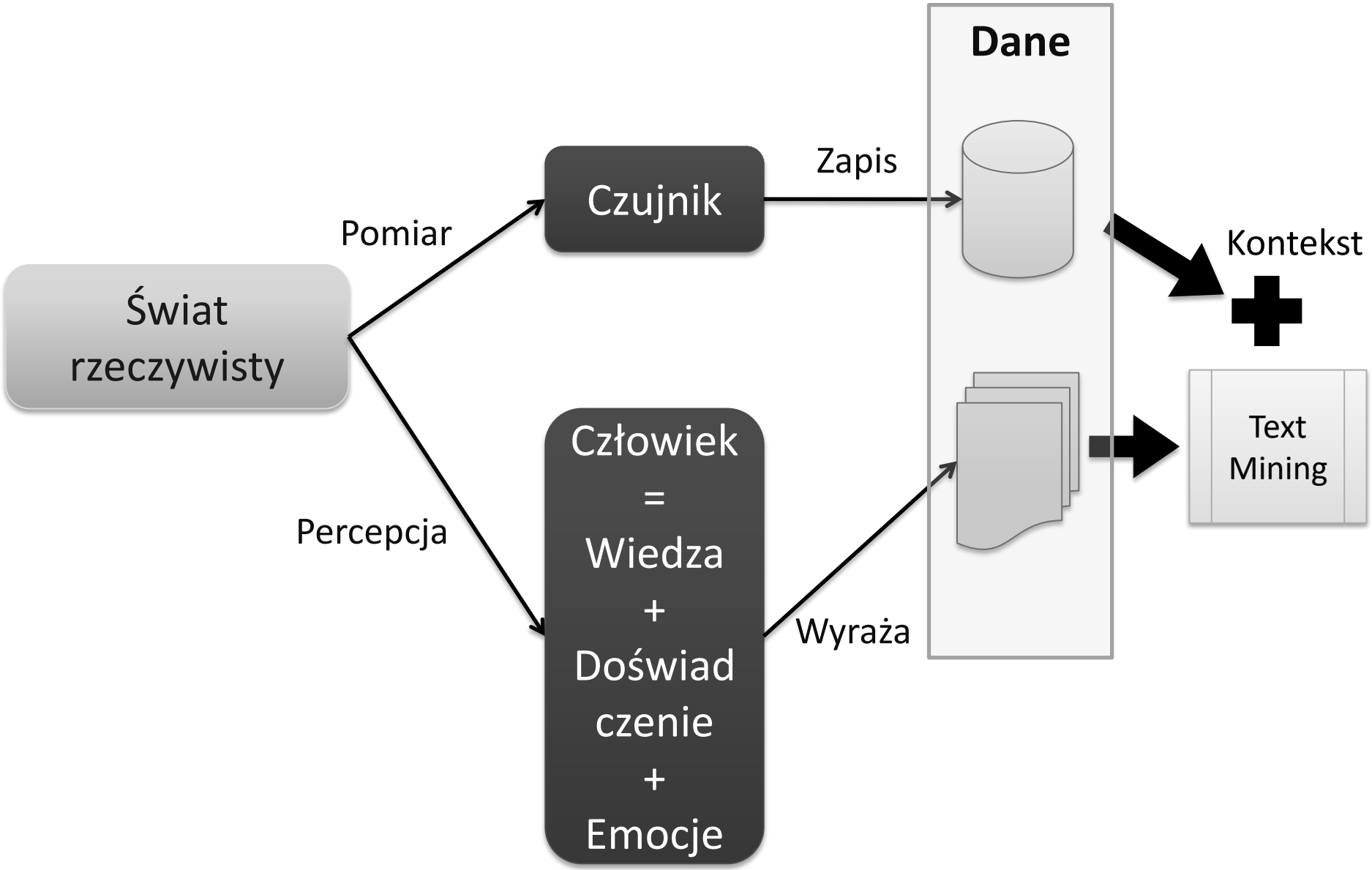

Pomiar
Świat rzeczywisty
Percepcja

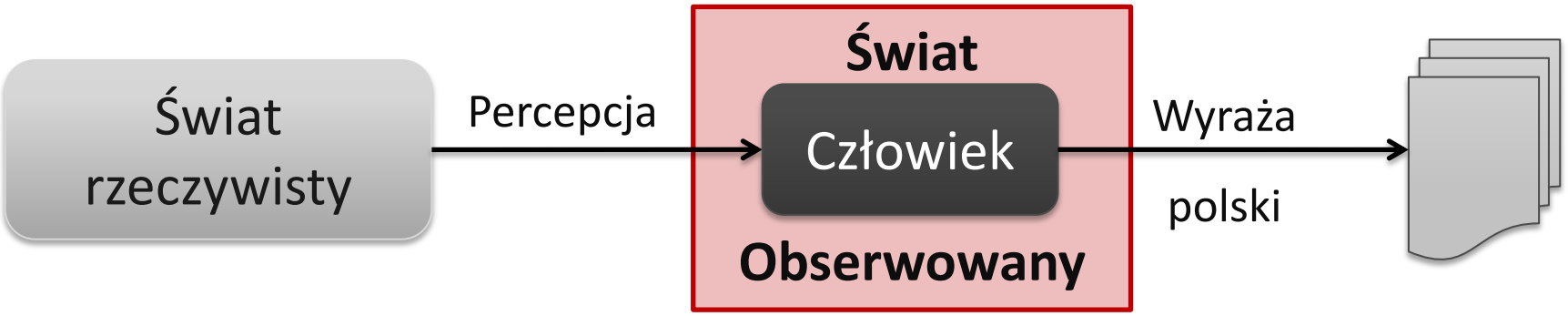

Dane
Zapis
Czujnik
Pomiar
Świat rzeczywisty
Percepcja
Człowiek
Wyraża

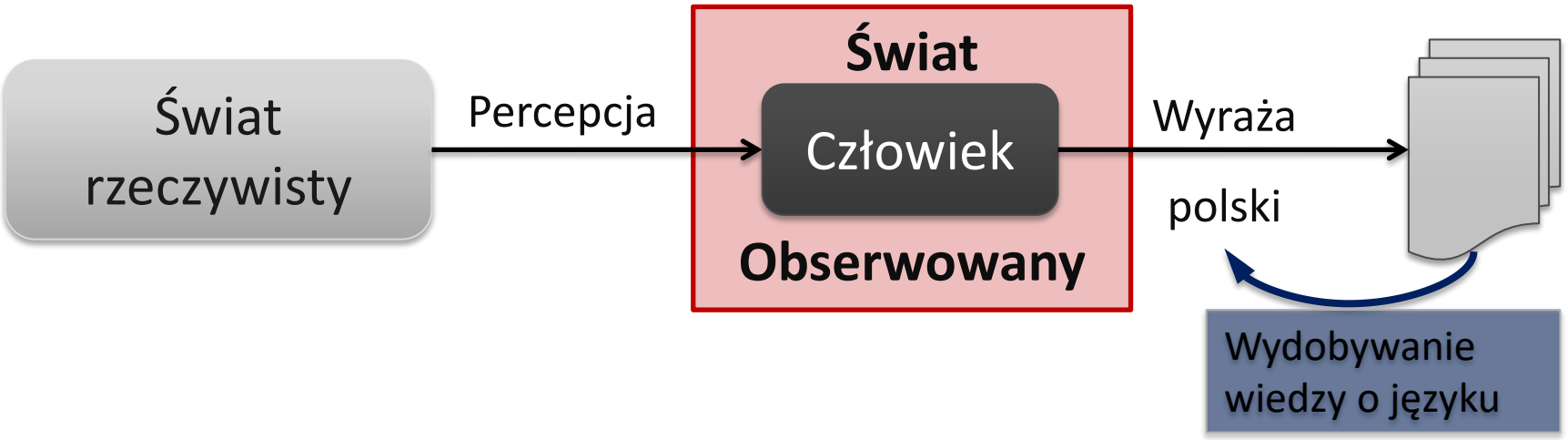

Dane
Świat rzeczywisty
Pomiar
Percepcja
Czujnik
Człowiek
Zapis
Wyraża
Data Mining
Text Mining

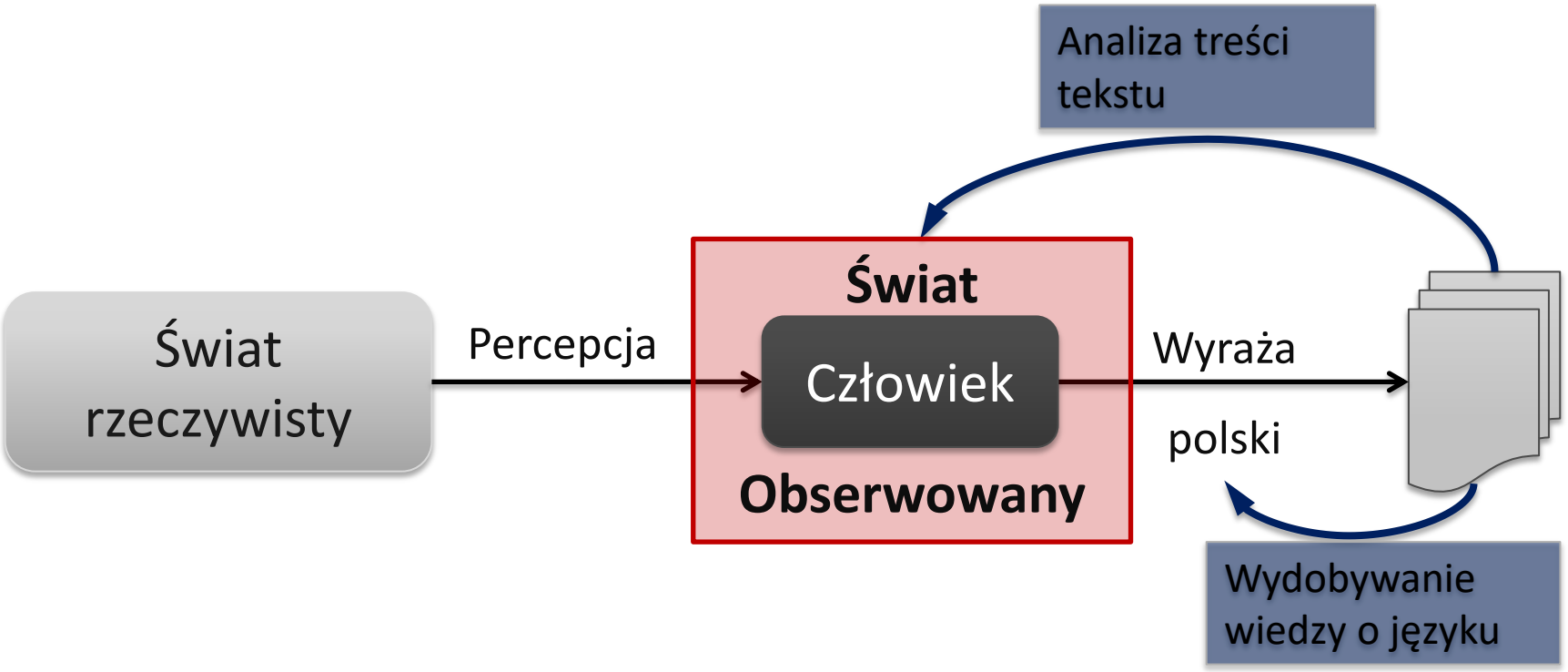

Dane
Świat rzeczywisty
Pomiar
Percepcja
Czujnik
Człowiek
= ?
Zapis
Wyraża
Data Mining
Text Mining

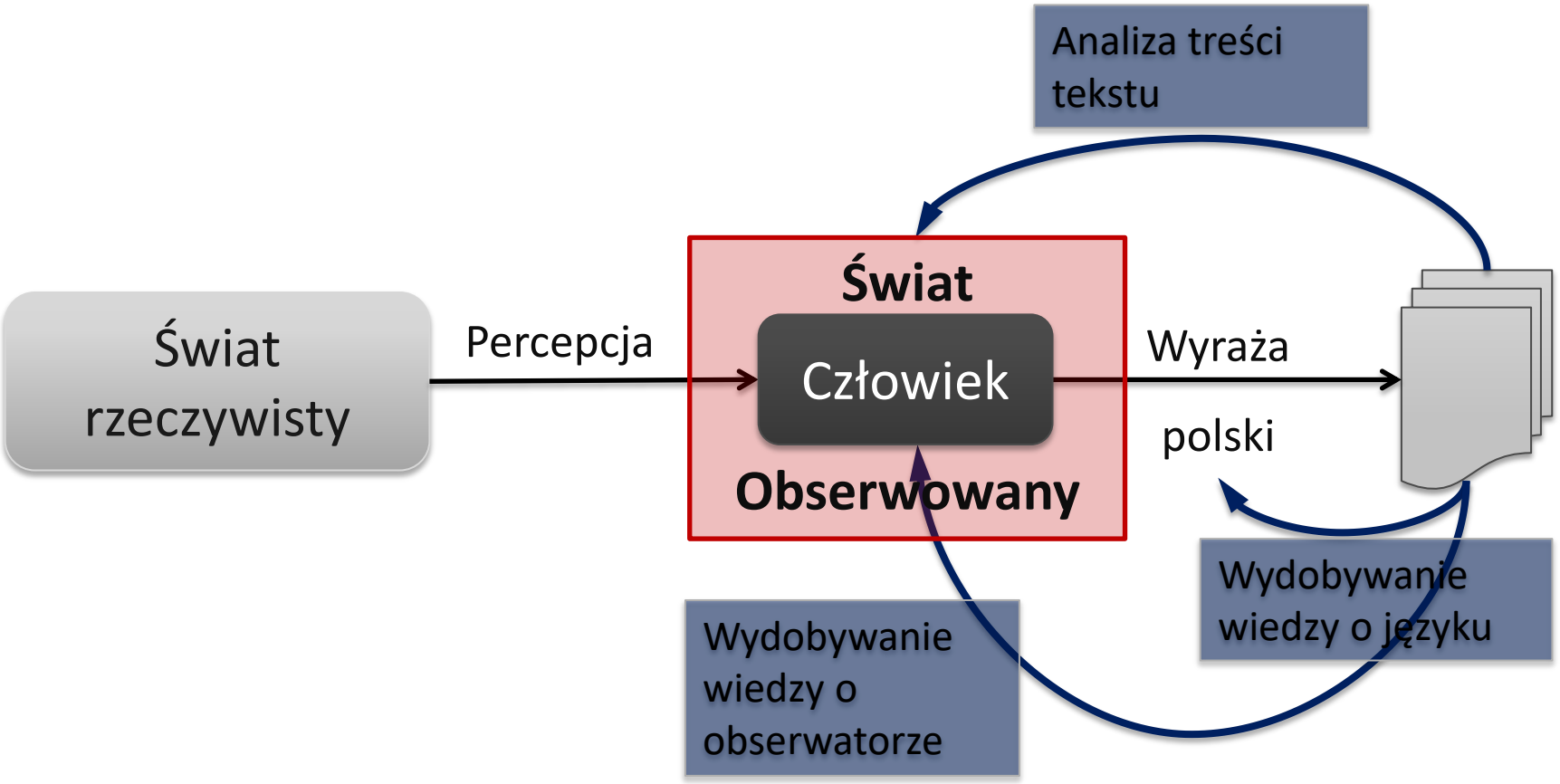

Dane
Świat rzeczywisty
Pomiar
Czujnik
Zapis
Data Mining
Percepcja
Człowiek = Wiedza + Doświadczenie + Emocje
Wyraża
Text Mining

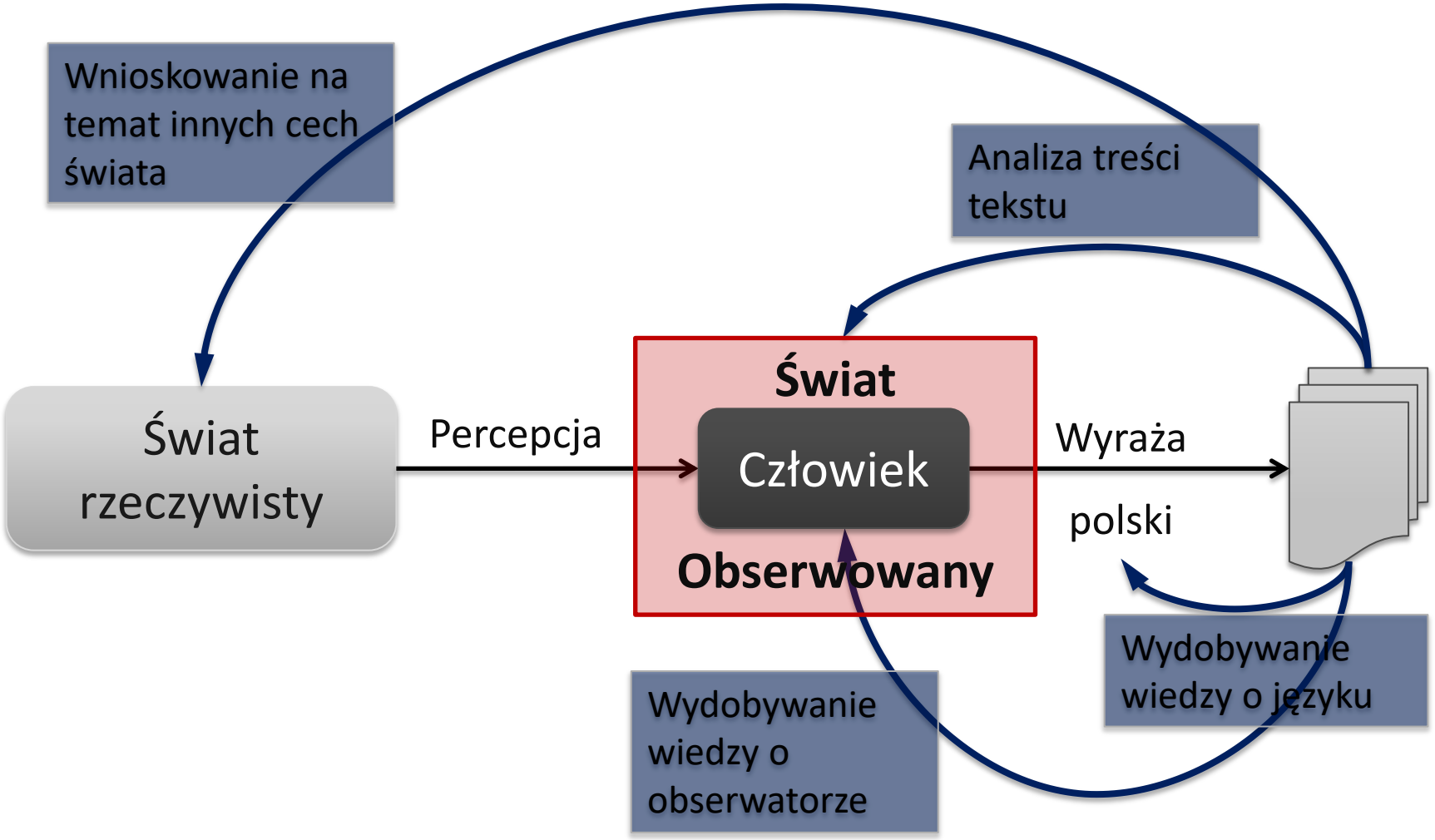

Dane
Świat rzeczywisty
Pomiar
Czujnik
Zapis
Kontekst
Percepcja
Człowiek = Wiedza + Doświadczenie + Emocje
Wyraża
Text Mining

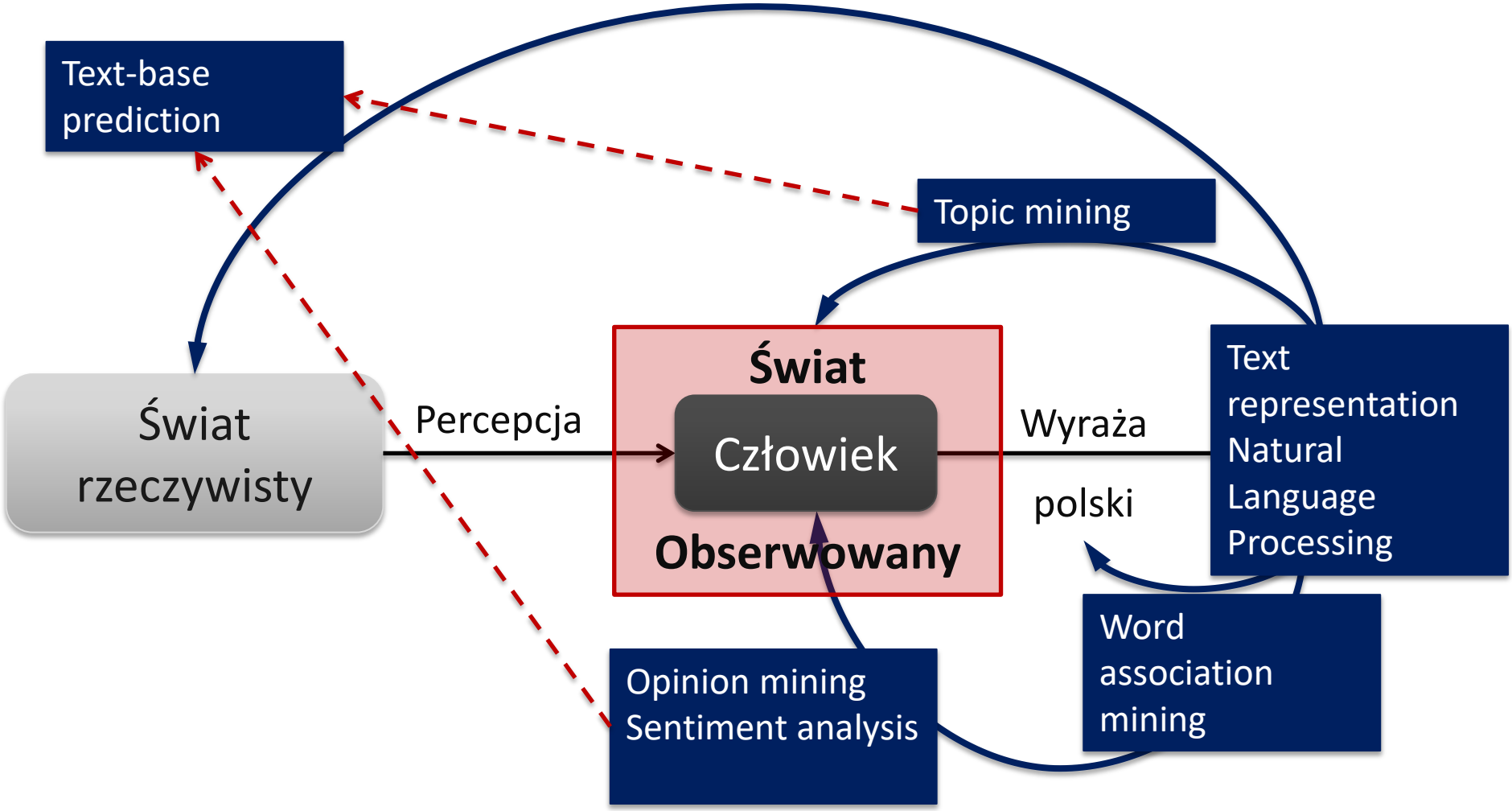

Świat
rzeczywisty
Percepcja
Świat
Człowiek
Obserwowany
Wyraża
polski


Świat
rzeczywisty
Percepcja
Świat
Człowiek
Obserwowany
Wyraża
polski
Wydobywanie
wiedzy o języku


Analiza treści tekstu
Świat
Człowiek
Obserwowany
Świat rzeczywisty
Percepcja
Wyraża
polski
Wydobywanie wiedzy o języku


Analiza treści tekstu
Świat
Człowiek
Obserwowany
Świat rzeczywisty
Percepcja
Wyraża
polski
Wydobywanie wiedzy o języku
Wydobywanie wiedzy o obserwatorze


Wnioskowanie na temat innych cech świata
Analiza treści tekstu
Świat rzeczywisty
Percepcja
Świat
Człowiek
Obserwowany
Wyraża
polski
Wydobywanie wiedzy o języku
Wydobywanie wiedzy o obserwatorze


Text-base prediction
Topic mining
Świat rzeczywisty
Percepcja
Świat
Człowiek
Obserwowany
Wyraża
polski
Text representation Natural Language Processing
Word association mining
Opinion mining Sentiment analysis

Odkrywanie wiedzy z tekstu – procedura

Uwzględniając aspekt analityczny text mining składa się z dwóch etapów:

- przekształcenie dokumentów źródłowych do postaci odpowiedniej do dalszej analizy – **przetwarzania języka naturalnego**,
- przeprowadzenie obliczeń i analiz pozwalających na osiągnięcie założonych celów **text mining**.

Proces Analizy Tekstu

Stworzenie korpusu

- Rezultatem etapu jest kolekcja dokumentów

Stworzenie macierzy term-dokument

- Rezultatem etapu jest „płaski” plik

Odkrycie wiedzy

- Rezultatem etapu są modele DM i wizualizacje

Proces Analizy Tekstu

Stworzenie korpusu

- Rezultatem etapu jest kolekcja dokumentów

Stworzenie macierzy term-dokument

- Rezultatem etapu jest „płaski” plik

Odkrycie wiedzy

- Rezultatem etapu są modele DM i wizualizacje

Proces Analizy Tekstu

Stworzenie korpusu

- Rezultatem etapu jest kolekcja dokumentów

Pre-processing

Stworzenie macierzy term-dokument

- Rezultatem etapu jest „płaski” plik

Odkrycie wiedzy

- Rezultatem etapu są modele DM i wizualizacje

KORPUSY

Korpus

KORPUS – zbiór tekstów reprezentatywnych dla języka, zapisany w formie elektronicznej, o ile to możliwe zawierający metadane

- niezbilansowany – niereprezentatywny dla języka, np. zawierający jedynie teksty o pewnej tematyce, albo też
- zbilansowany - reprezentatywny dla całego języka
- jednojęzykowy vs. wielojęzyczny (bitext)
- anotowany – zawierający metadane, w szczególności POS tags i/lub informacje o rozbiorze zdania

korpus jest zwykle statyczny i jako taki jest „fotografią” języka w pewnej chwili – np. *Brown corpus* – język angielski z lat 60-tych

Korpus

http://sjp.pwn.pl/korpus (70 mln słów)

dostępnych za darmo, ale bez możliwości pobrania

http://nkjp.pl/ - korpus IPI PAN

http://zil.ipipan.waw.pl/ - strona The Linguistic Engineering Group (PAN)

Korpus – text mining

Korpus to dowolny zbiór tekstów, w którym czegoś szukamy.
O korpusach w tym znaczeniu mówią najczęściej językoznawcy, ale także archiwiści, historycy i informatycy.

- Media społecznościowe/Komentarze/Blogi
- Artykuły
- Książki
- Dokumenty prawnicze
- Dokumenty wytwarzane przez pracowników/CV

https://www.washingtonpost.com/news/the-fix/wp/2016/09/26/the-first-trump-clinton-presidential-debate-transcript-annotated/

Korpus – text mining

- Stworzenie korpusu
  - Zebranie wszystkich relewantnych nieustrukturyzowanych danych (np. dokumenty tekstowe, pliki XML, emaile, strony Web, notatki, nagrania głosowe)
  - Zamiana na postać cyfrową i ustandaryzowanie zbioru (np. wszystko jako pliki tekstowe ASCII)
  - Zgromadzenie dokumentów we wspólnym miejscu (np. w pliku płaskim, lub w folderze jako osobne pliki)

PRZETWARZANIA JĘZYKA NATURALNEGO

Analiza języka naturalnego

NLP (*Natural Language Processing*)

zbiór technik komputerowych służących do analizy i reprezentacji tekstów występujących na poziomie analizy lingwistycznej w celu uzyskania, przypominającego ludzki, sposobu przetwarzania języka w określonym zakresie zadań i zastosowań

NLP, NLU, NLG

NLP – Natural Language Processing

- Inne nazwy: Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

NLU – Natural Language Understanding

- Dosłownie „rozumienie języka naturalnego” - semantyka i logika

NLG – Natural Language Generation

Język naturalny

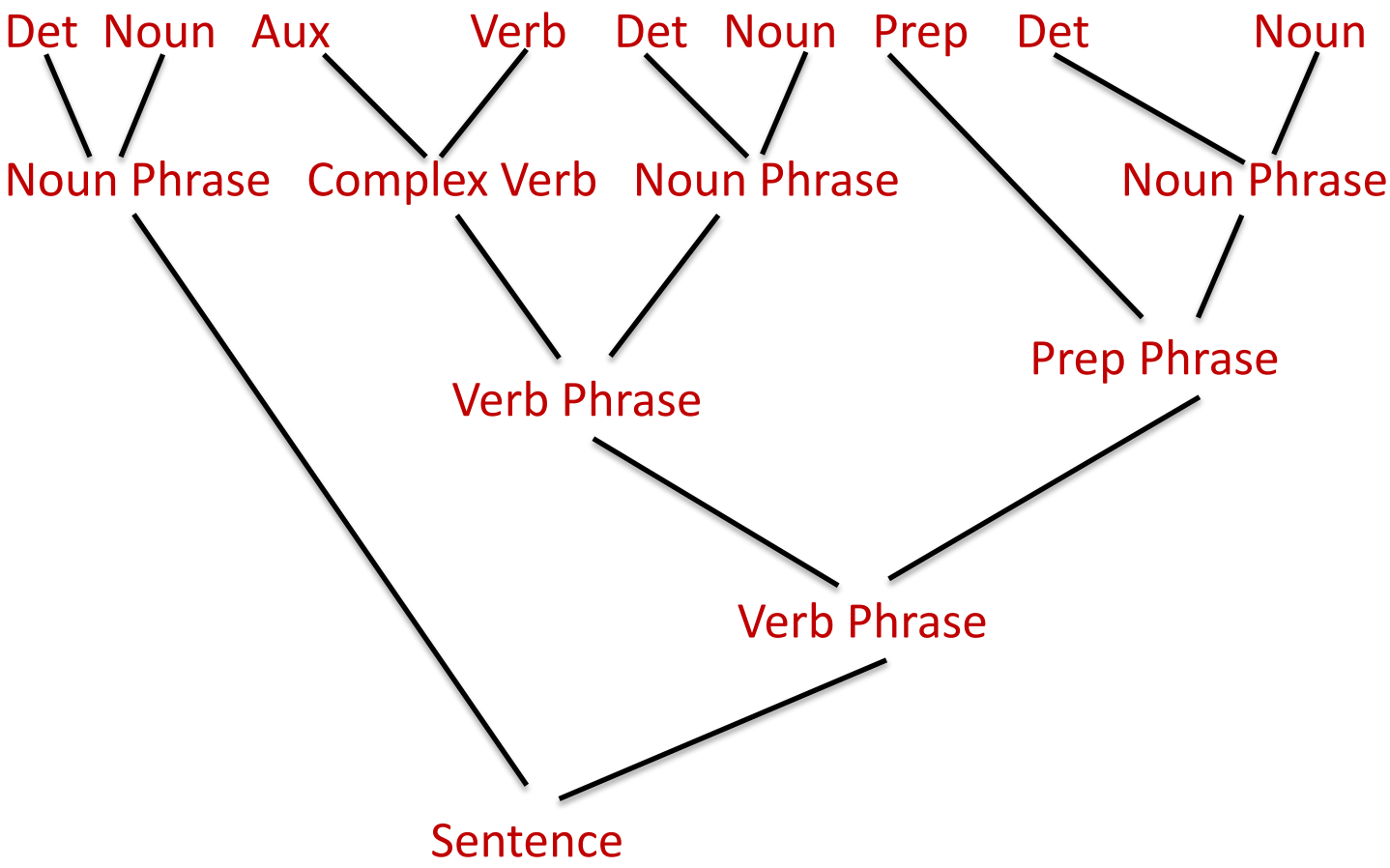
- Nieprecyzyjny (na wszystkich poziomach)
  - Fonetyka, morfologia: dźwięki i słowa
  - Składnia: zdania i ich struktura
  - Semantyka: znaczenie treści wypowiedzi
  - Pragmatyka: znaczenie samej wypowiedzi „w świecie”
- Skomplikowany (nawet jeśli uznać reguły gramatyczne)
- Wymaga posiadania wiedzy o świecie

A dog is barking a boy on the playground

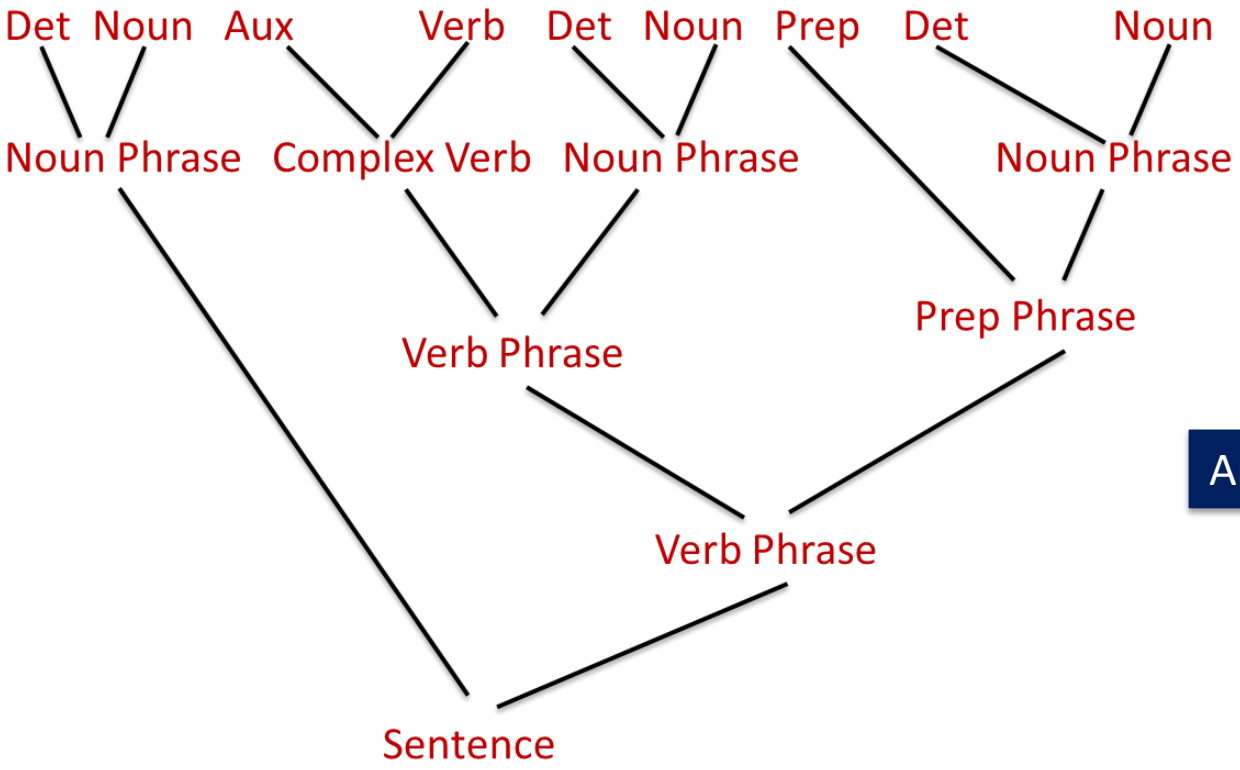
A dog is barking a boy on the playground

Det Noun Aux Verb Det Noun Prep Det Noun

A dog is barking a boy on the playground

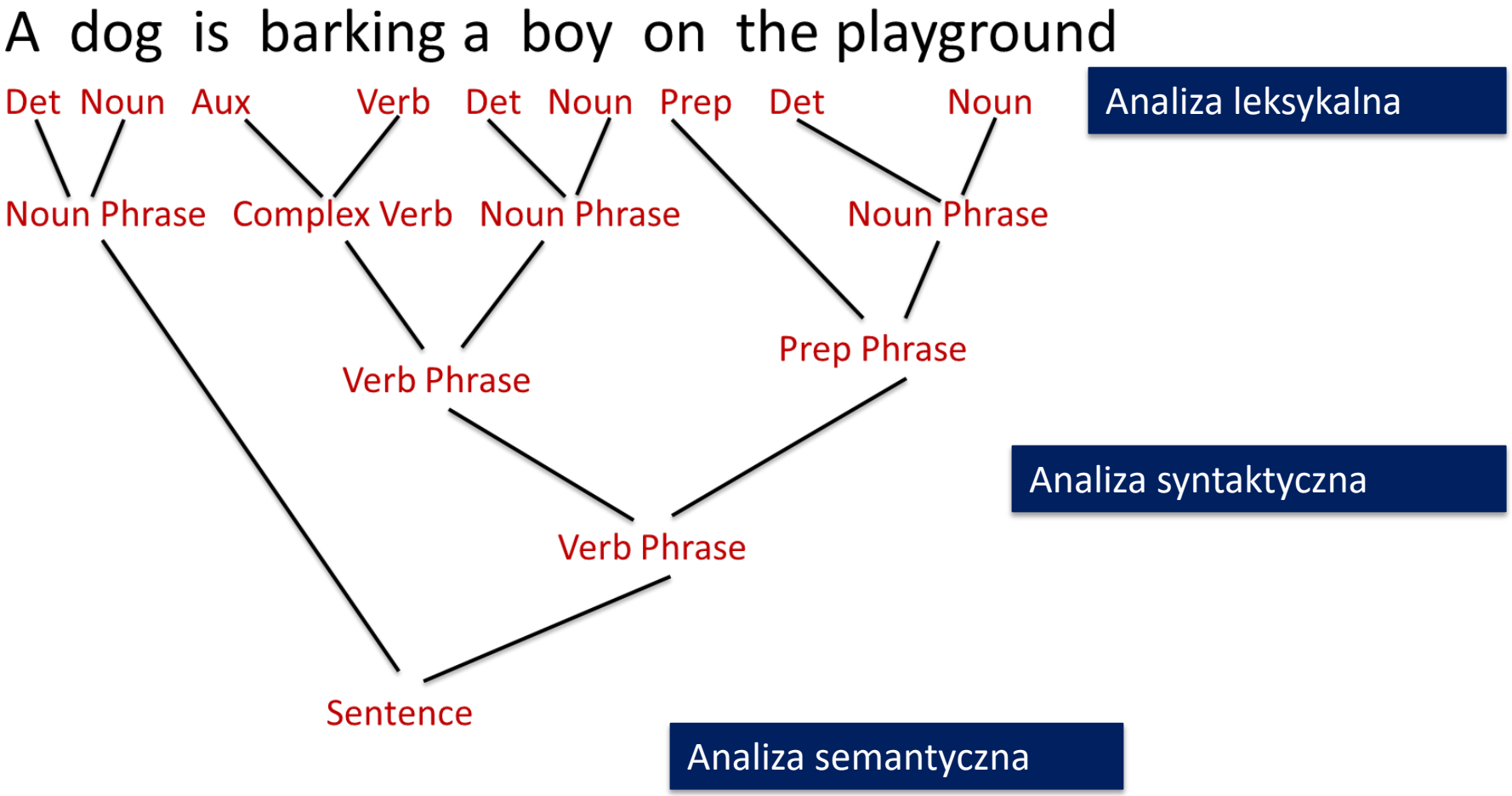


A dog is barking a boy on the playground



Analiza leksykalna

Analiza syntaktyczna

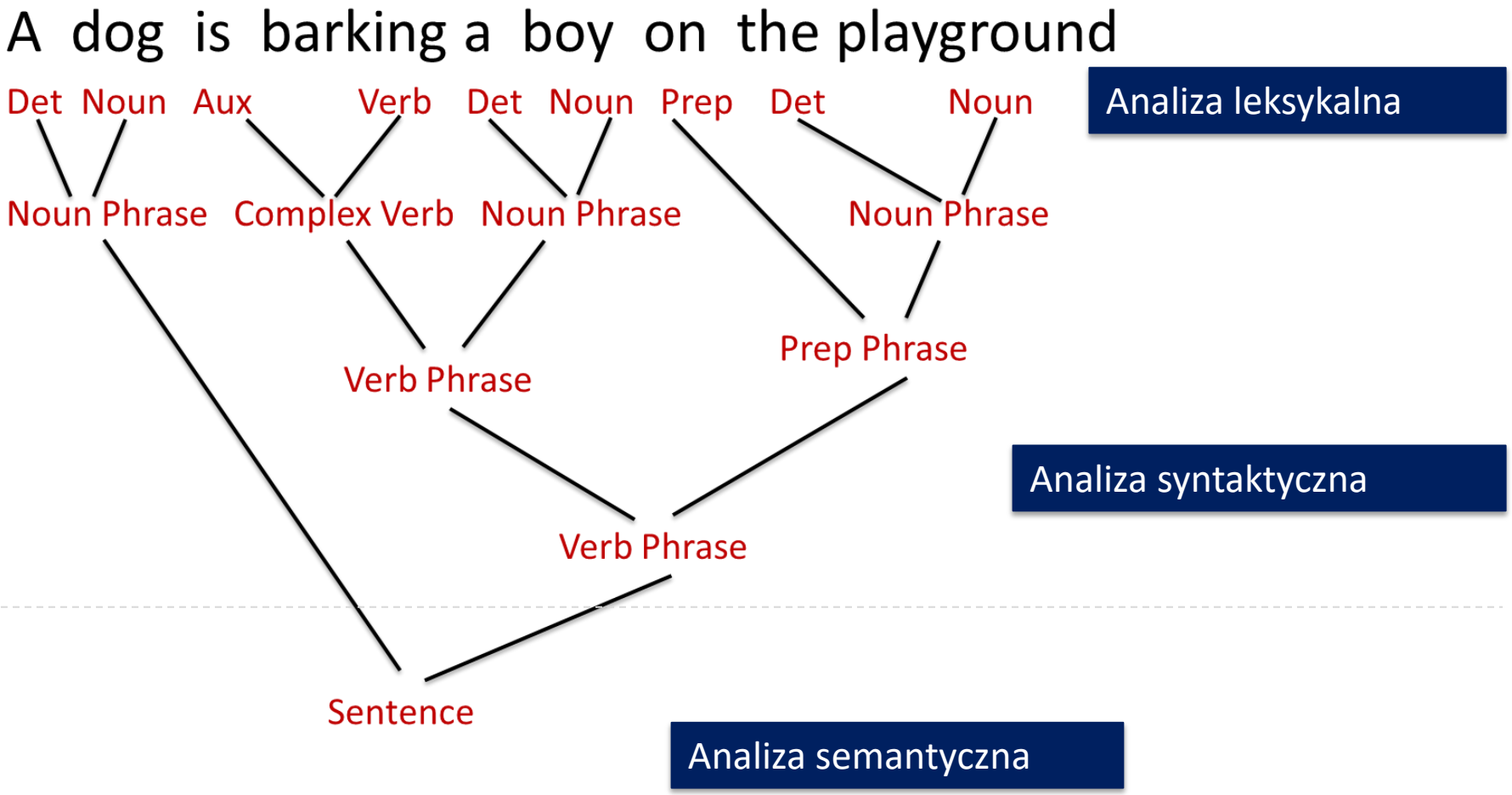

A dog is barking a boy on the playground
Det Noun Aux Verb Det Noun Prep Det Noun
Noun Phrase Complex Verb Noun Phrase Noun Phrase
Verb Phrase
Prep Phrase
Verb Phrase
Sentence
Analiza leksykalna
Analiza syntaktyczna
Analiza semantyczna

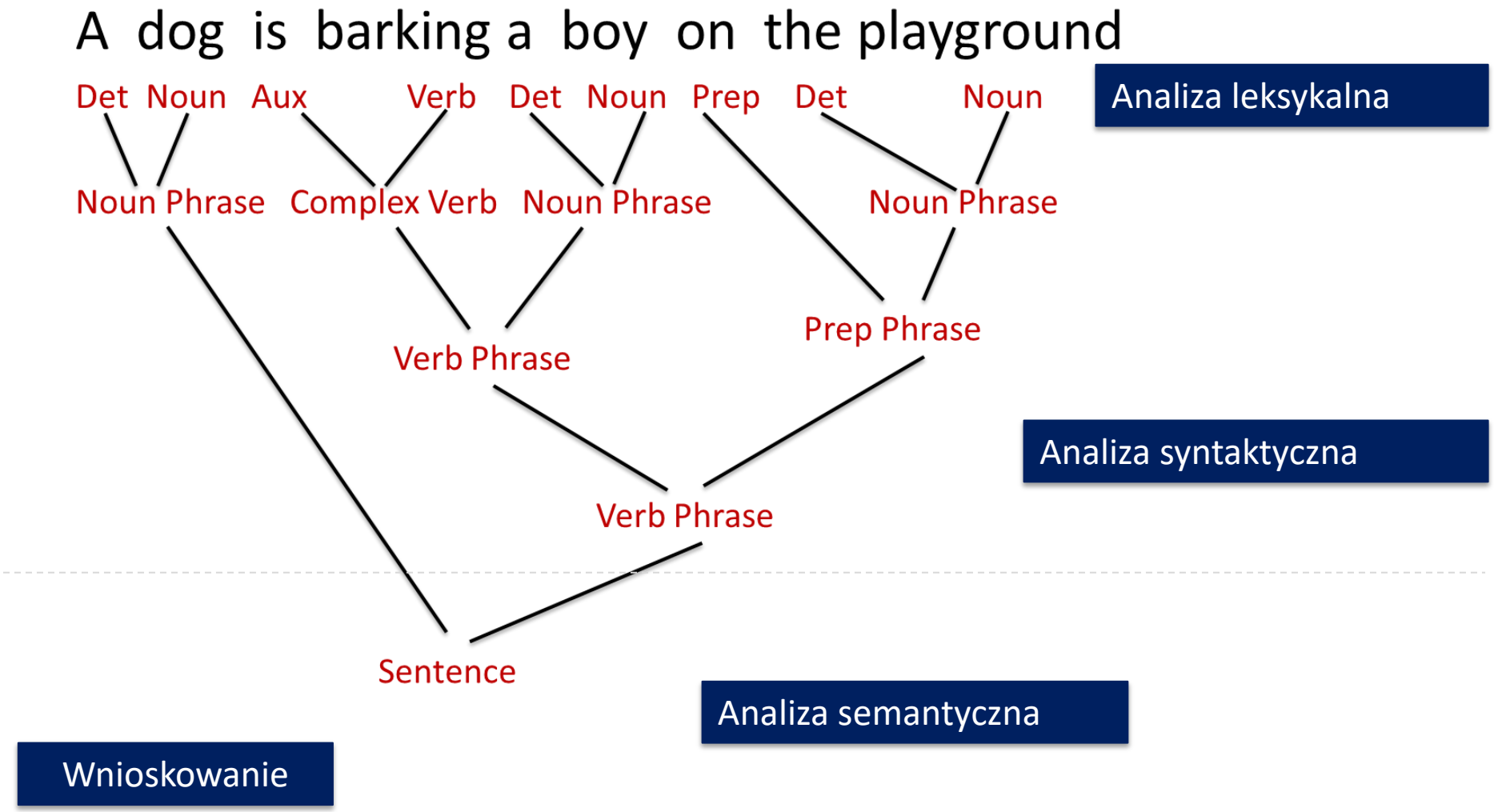
Syntaktyka a semantyka

- Plac na pies zabaw na chłopca szczeka

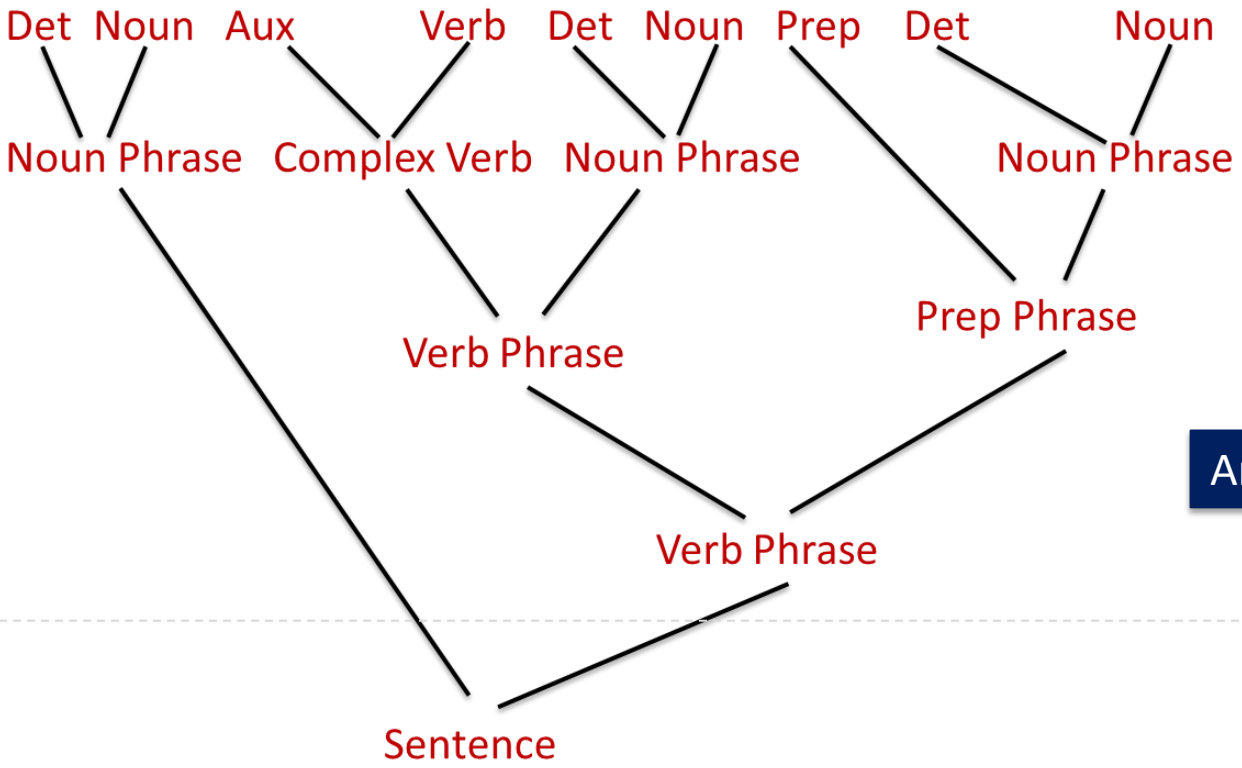
Syntaktyka a semantyka

- *Plac na pies zabaw na chłopca szczeka*
- *Bezbarwne zielone idee wściekle śpią* (*Colorless green ideas sleep furiously*)


A dog is barking a boy on the playground
Det Noun Aux Verb Det Noun Prep Det Noun
Noun Phrase Complex Verb Noun Phrase Noun Phrase
Verb Phrase
Prep Phrase
Verb Phrase
Sentence
Analiza leksykalna
Analiza syntaktyczna
Analiza semantyczna


A dog is barking a boy on the playground
Det Noun Aux Verb Det Noun Prep Det Noun
Noun Phrase Complex Verb Noun Phrase Noun Phrase
Verb Phrase
Prep Phrase
Verb Phrase
Sentence
Analiza leksykalna
Analiza syntaktyczna
Analiza semantyczna
Wnioskowanie

A dog is barking a boy on the playground



Analiza leksykalna

Analiza syntaktyczna

Analiza semantyczna

Wnioskowanie

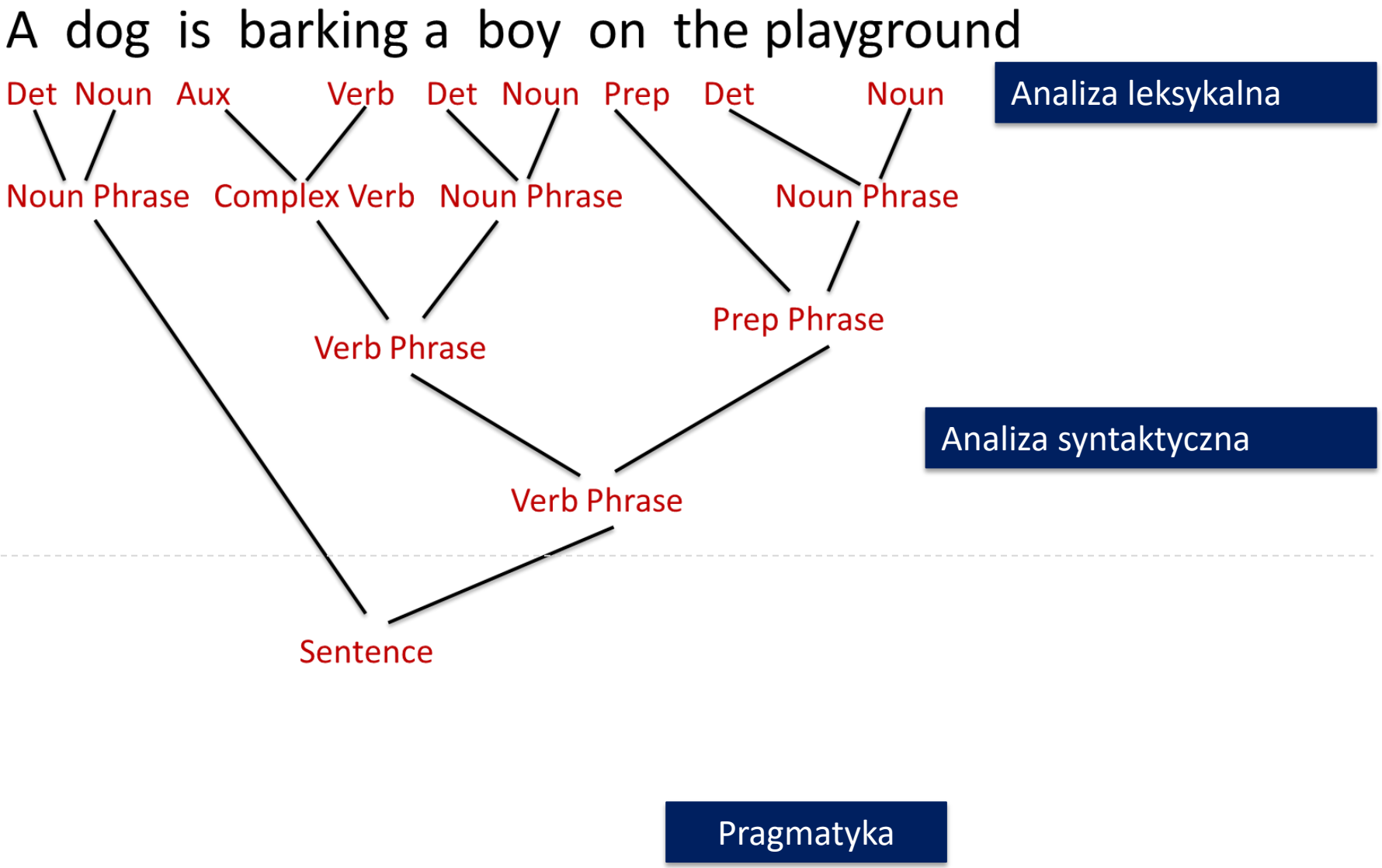
Dog(d1).
Boy(b1).
Playground(p1).
Barking(d1,b1,p1).

+

Scared(x) if
Barking(_,x,_).

=

Scared(b1)


A dog is barking a boy on the playground
Det Noun Aux Verb Det Noun Prep Det Noun
Noun Phrase Complex Verb Noun Phrase Noun Phrase
Verb Phrase
Prep Phrase
Verb Phrase
Sentence
Analiza leksykalna
Analiza syntaktyczna
Pragmatyka

Język polski

Język polski

Pies szczeka na chłopca na placu zabaw

Język polski

Pies szczeka na chłopca na placu zabaw

Pies oszczekuje chłopca na placu zabaw

Pies obszczekuje chłopca na placu zabaw

Język polski

Pies szczeka na chłopca na placu zabaw

Pies oszczekuje chłopca na placu zabaw

Pies obszczekuje chłopca na placu zabaw

Na placu zabaw Pies szczeka na chłopca

Język polski

Fleksja

Pies szczeka na chłopca na placu zabaw

Pies oszczekuje chłopca na placu zabaw

Pies obszczekuje chłopca na placu zabaw

Dowolny szyk zdania

Na placu zabaw Pies szczeka na chłopca

Podejścia do analizy języka naturalnego

Podejścia do analizy języka naturalnego

„Gramatyczne”

- Język naturalny można opisać wykorzystując aparat logiki matematycznej
- Lingwistyka porównawcza – Jakob Grimm, Rasmus Rask
- Noam Chomsky – I-Language i E-language

Podejścia do analizy języka naturalnego

„Statystyczne”

- Przekonanie, iż struktura i reguły użycia słów w języku naturalnym można odkryć, analizując rzeczywiste wypowiedzi
- Najlepiej analizować dużo (bardzo dużo) wypowiedzi…

Podejścia do analizy języka naturalnego?

Aby skutecznie analizować język naturalny konieczne jest wykorzystywania zarówno statystyki jak i podejścia „gramatycznego”

- „czyste” modele oparte na regułach można wzbogacić o elementy probabilistyczne;
- metody statystyczne można wzbogacić poprzez wykorzystanie znanych reguł i źródeł „twardej” wiedzy (np. word sense disambiguation + słowniki + ontologie)

Statystyczna

- Problem dwuznaczności słów

Statystyczna

- Problem dwuznaczności słów

  Znaleźli perłę w muszli.

Statystyczna

- Problem dwuznaczności słów

Znaleźli perłę w **muszli**.

muszla małża, muszla klozetowa

Statystyczna

- Problem dwuznaczności słów

Znaleźli perłę w **muszli**.

Dysponując korpusem poprawnych tekstów w języku polskim obliczyć można prawdopodobieństwo:

P_1 – współwystępowanie <perła, małż>

P_2 – współwystępowanie <perła, klozet>

Zadania NLP

- Wyodrębnianie zdań
- Wyodrębnianie słów – tokenizacja
- Zliczanie wystąpień słów (unikatowych, powtarzających się)
- Częstotliwość wystąpień słów
- Normalizacja
- Stemming, Lematyzacja

Zadania NLP

- Parsowanie zdania
- Rozbiór gramatyczny zdania (**POS** – *Part of Speech tagging*)
- Identyfikacja semantyczna
- Rozpoznawanie jednostek nazwanych (**NER** – *Named Entity Recognition*)

PROCES PRZYGOTOWANIA TEKSTU DO ANALIZY

Proces przygotowania surowego tekstu do analizy

1. Segmentacja – podzielenie wejściowego ciągu znaków (dokumentu) na sekcje, zdania lub związki frazeologiczne, wyrazy.

Proces przygotowania surowego tekstu do analizy

1. Segmentacja – podzielenie wejściowego ciągu znaków (dokumentu) na sekcje, zdania lub związki frazeologiczne, wyrazy.

   Gdy celem analizy jest wyodrębnienie z tekstu określonych informacji w procesie segmentacji rozróżniać można ciagi specjalistyczne np.: daty, liczby, adresy itp..

Proces przygotowania surowego tekstu do analizy

1. Tokenizacja – to proces w wyniku którego monolityczny tekst zostaje podzielony na ciąg pojedynczych tokenów.

Token to ciąg znaków ograniczony ustalonymi separatorami takimi jak spacja czy przecinek (separatorem może być dowolny ciąg który można opisać w postaci wyrażenia regularnego).

http://lucene.apache.org/

http://nlp.stanford.edu/software/tokenizer.shtml

Tokenizacja – różne podejścia

A dog is barking a boy on the playground ->

A; dog; is; barking; a; boy; on; the; playground;

Tokenizacja – różne podejścia

A dog is barking a boy on the playground ->

A; dog; is; barking; a; boy; on; the; playground;

pop-art ->

pop; art; / pop-art;

A.M. ->

A.; M.; / A.M.;

Proces przygotowania surowego tekstu do analizy

2. POS tagging – rozpoznanie **części mowy** każdego z tokenów wyodrębnionych w tekście.
   - informacja o częściach mowy pozwala na wzbogacenie zbioru cech, w celu rozróżnienia sensu znaczeniowego danej wypowiedzi.

http://nlp.stanford.edu/software/tagger.shtml

http://morfologik.blogspot.com/

http://nlp.pwr.wroc.pl

POS tagging – j. angielski

PART-OF-SPEECH	TAG	PRZYKŁAD
Adjective	JJ	happy, bad
Adjective, comparative	JJR	happier, worse
Adjective, cardinal number	CD	3, fifteen
Adverb	RB	often, particularly
Conjunction, coordination	CC	and, or
Conjunction, subordinating	IN	although, when
Determiner	DT	this, each, other, the, a, some
Determiner, postdeterminer	JJ	many, same
Noun	NN	aircraft, data
Noun,plural	NNS	women, books
Noun, proper, singular	NNP	London, Michael
Noun, proper, plural	NNPS	Australians, Methodists
Pronoun, personal	PRP	you, we, she, it
Pronoun, question	WP	who, whoever
Verb, base present form	VBP	take, live

POS tagging – j. polski

- Morfologik – README.Polish.txt

Proces przygotowania surowego tekstu do analizy

3. Stop-words – lista wyrazów nieniosących ze sobą istotnych treści
   - dobór wyrazów jest sprawą arbitralną, listy mogą zawierać różną liczbę wyrazów,
   - etap może być także pomięty.

Proces przygotowania surowego tekstu do analizy

4. Stemming – tokeny (formy wyrazowe) zastępowane są swoimi niezmienialnymi częściami. W procesie tym następuje ekstrakcja **rdzenia** (*stem*).

Stemming

Rdzeń to podstawowy morfem wyrazu, do którego mogą być dołączane morfemy afiksalne (afiksy), zarówno fleksyjne, jak i słowotwórcze. Rdzeń jest **morfemem leksykalnym**.

Morfem najmniejsza znacząca cząstka wyrazu.

Afiksy to morfemy związane, niosące funkcje gramatyczne. Dzielą się na przedrostki, przyrostki i wrostki.

Funkcjonalna klasyfikacja morfemów

Morfem leksykalny

Przykłady: **kot**-k-owi, za-**smuć**-i-ł się.

Morfem gramatyczny

Przykłady: kot-k-**owi**, za-smuć-**i**-**ł** się.

Morfem słowotwórczy

Przykłady: kot-**k**-owi, **za**-smuć-i-ł się.

Stemming – algorytm Portera

Polega na iteracyjnym usuwaniu nadmiarowych sufiksów.

Algorytm składa się z 5 głównych kroków.

http://snowballstem.org/

Stemming – algorytm Portera

Algorytm:

1. Depluralyzacja oraz proste końcówki (usuwanie -es, -ed, -ing etc.)
2. Redukcja podwójnych sufiksów (“ational”→“ate”, “tional”→“tion”, etc.)
3. Usuwanie form przysłówkowych, bezokolicznikowych i im podobnych (“ness”→“ ”, “alize”→“al”, “icate”→“ic”)
4. Usuwanie “ant”, “ence” etc.
5. Usuwanie końcówki “e” oraz redukcja podwójnych spółgłosek (“ll”→“l”)

Proces przygotowania surowego tekstu do analizy

4. Lematyzacja – tokeny (formy wyrazowe) zastępowane są formami podstawowymi - **lemmami**.

Proces przygotowania surowego tekstu do analizy

4. Lematyzacja – tokeny (formy wyrazowe) zastępowane są formami podstawowymi - lemmami.

   W językach z bogatą i nieregularną fleksją stemming jest praktycznie niemożliwy, w takich przypadkach należy zastosować lematyzację.

Lematyzacja

Lemma to kanoniczna forma leksemu, która używana jest do jego reprezentacji.

Lematyzacja

Lemma to kanoniczna forma leksemu, która używana jest do jego reprezentacji.

Leksem to wyraz słownikowy. Jednostka systemu słownikowego języka, na którą składa się:

- znaczenie leksykalne,
- zespół wszystkich funkcji gramatycznych, jakie dany leksem może spełniać,
- zespół form językowych reprezentujących w tekście leksem w jego poszczególnych funkcjach.

kotek

Słownik języka polskiego pod red. W. Doroszewskiego

kotek *mIII, D.* kotka zdr. od kot 1. w zn. 1: Skoczył mi na kolana kotek z podniesionym do góry ogonkiem, łasząc się i mrucząc. Unił. *Pok. 193.* Ta twoja pani, jak rozumiem, lubi grę w kotka i myszkę, tylko ona chce być kotem. Weys. Józ. *Puszcza 150.* Wlazł kotek — na płotek i mruga: piękna to piosneczka niedługa. Kolb. *Pieśni 448.* Kotek się ciągle myje, będziem gościa mieli. Groza *Poezje 141.*
◊ *przen.* W tej chwili podchwyciła moje przerażone spojrzenie i powiedziała łagodnie: Kotku, idź do salonu. Kow. A. *Rogat. 67.* Były znów dwa kotki: Maniusia i Michalina... Ach! palą też, palą tymi ognistymi oczyma! Żer. *Dzien. I, 233.*
2. p. kot w zn. 4b: W szufladzie stolika znalazł doktor rękopis owej „Fizyki" (...) w szafce trochę bielizny, salopkę kotkami podbitą, jakąś starą czarną sukienkę. Żer. *Opow. II, 105.*
3. *żegl.* w *lm* «ochraniacze z rozkręconych pokrętek liny zaplecionych na stalówce dla ochrony żagli od przetarcia» // *L*

Lematyzacja

W języku polskim za lemmy poszczególnych części mowy służą:

- czasownik - forma bezokolicznika (mieszkać)
- rzeczownik - mianownik w liczbie pojedynczej (dom)
- przymiotnik - mianownik w liczbie pojedynczej rodzaju męskiego w stopniu równym (niski)
- przysłówek - stopień równy (wolno)
- liczebnik i zaimek - odpowiednie cechy części mowy, od których wywodzi się dany typ liczebnika/zaimka (mój, jeden)

Lematyzacja – j. polski

http://morfologik.blogspot.com/

Lematyzacja – j. polski

http://morfologik.blogspot.com/

forma podstawowa;
forma odmieniona;
znaczniki gramatyczne

Lematyzacja – j. polski

http://morfologik.blogspot.com/

podstawa;odmiana;znaczniki gramatyczne

kotek;kotek;subst:sg:nom:m2
kotka;kotek;subst:pl:gen:f

Lematyzacja – j. polski

kotek;kotek;subst:sg:nom:m2
kotka;kotek;subst:pl:gen:f

* subst - rzeczownik
* sg / pl - liczba pojedyncza / liczba mnoga
* nom - mianownik
* gen - dopełniacz
* m1, m2, m3 - rodzaje męskie
* f - rodzaj żeński

Lematyzacja – j. polski

kotek;kotki;
subst:pl:acc:m2
 +subst:pl:nom:m2
 +subst:pl:voc:m2
kotka;kotki;
subst:pl:acc:f
 +subst:pl:nom:f
 +subst:pl:voc:f
 +subst:sg:gen:f

Proces przygotowania surowego tekstu do analizy

5. Normalizacja tekstu - przetworzenie tekstu, nadający mu spójną formę, ułatwiającą dalszą interpretację.

Normalizacja

- zmiana wielkości liter (na małe lub wielkie),
- normalizacja skrótów,
- normalizacja wyrażeń numerycznych,
- normalizacja znaków specjalnych,
- zmiana lub usuwanie znaków interpunkcyjnych,
- usuwanie (lub zmienianie) znaków diakrytycznych.

Model wektorowy

REPREZENTACJA TEKSTU

Model wektorwy

- Każdy jednostkowy dokument zapisany w jezyku naturalnym reprezentowany jest za pomocą wektora.

Model wektorwy

- Każdy jednostkowy dokument zapisany w jezyku naturalnym reoprezentowany jest za pomocą wektora
- Każdy wektor składa się z liczb (w_{ij}) będących miarą istotności tokenów/stemów/lemm.

Model wektorwy

- Korpus, składający się *m* dokumentów przekształcony zostanie w zbiór *m* wektorów tworzących wielowymiarową przestrzeń.

Model wektorwy

- Korpus, składający się *m* dokumentów przekształcony zostanie w zbiór *m* wektorów tworzących wielowymiarową przestrzeń.
- Macierz term-dokument

Model wektorwy

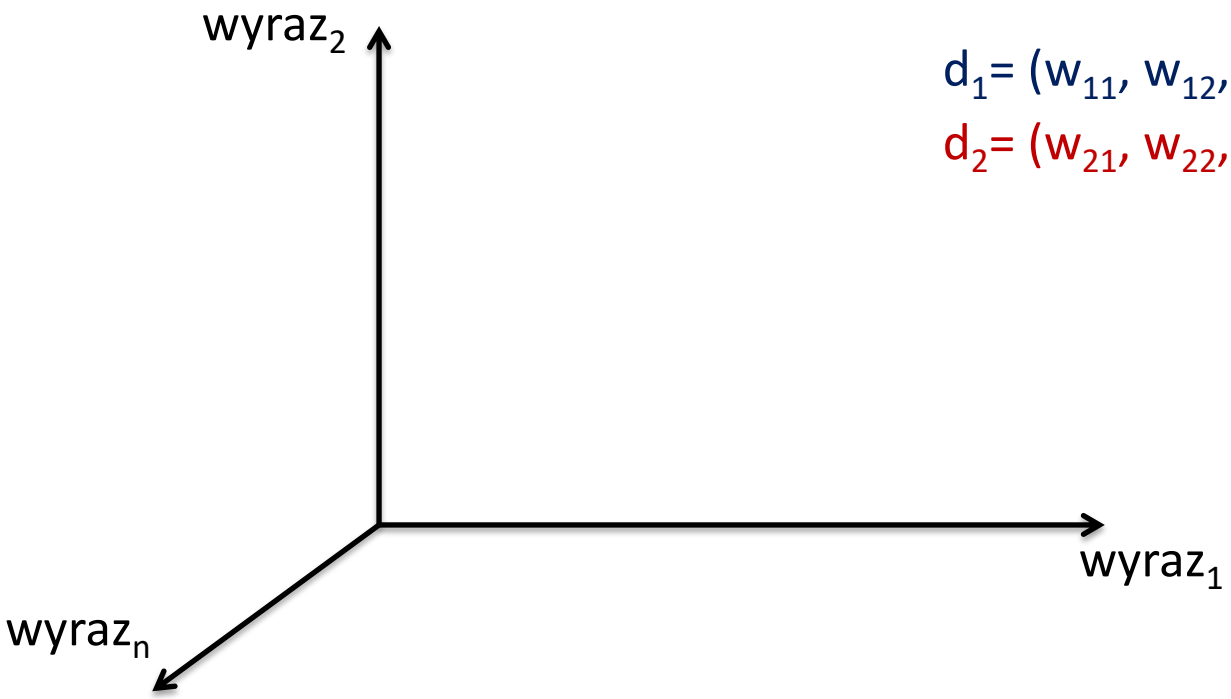
- BoW - *Bag of Words*
- SVM - *Vector Space Model*

Vector Space Model

$d_1 = (w_{11}, w_{12}, \ldots, w_{1N})$

$d_2 = (w_{21}, w_{22}, \ldots, w_{2N})$

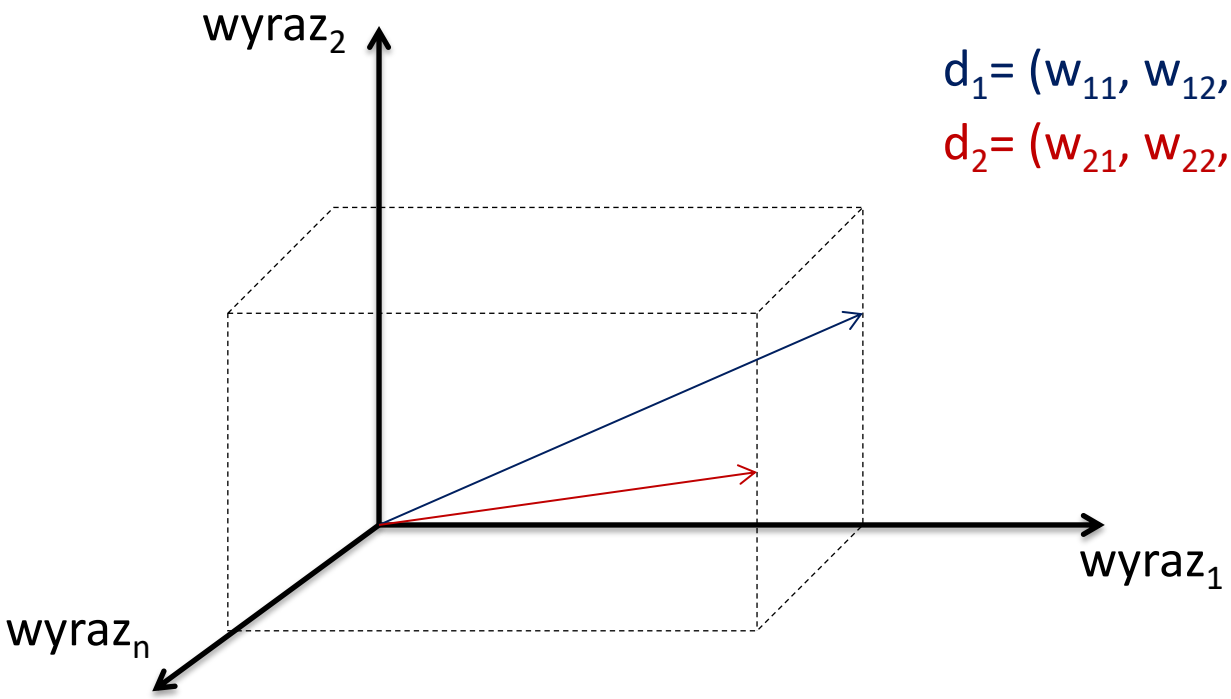
Vector Space Model



$d_1 = (w_{11}, w_{12}, \ldots, w_{1N})$

$d_2 = (w_{21}, w_{22}, \ldots, w_{2N})$

Vector Space Model



$d_1 = (w_{11}, w_{12}, \ldots, w_{1N})$

$d_2 = (w_{21}, w_{22}, \ldots, w_{2N})$

Macierz Term-Dokument

Macierz Term-Dokument

- macierz wystąpień – powszechny schemat reprezentacji związków między termami a dokumentem, w postaci tabelarycznej, gdzie w wierszach (kolumnach) są termy, w kolumnach (wierszach) dokumenty, a w komórkach częstotliwość występowania **termów** w dokumentach

term – słowo lub kilka słów (związki frazeologiczne) wyekstrahowane bezpośrednio z korpusu za pomocą metod przetwarzania języka naturalnego (NLP)

Macierz Term-Dokument

Dokument / Term	Dok1	Dok2	Dok3	Dok4	…
cooffee	3	1	2	0	
cup	0	1	1	0	
like	2	0	0	1	
shop	0	0	1	0	

Macierz Dokument-Term

- Czasami przydatna jest transponowana macierz Term-Dokument

Macierz Dokument-Term

Term / Dokument	cooffee	cup	like	shop	…
Dok1	3	0	2	0	
Dok2	1	1	0	0	
Dok3	2	1	0	1	
Dok4	0	0	1	0	

MIARY WAŻNOŚCI

Miary ważności

- *frequency weight* - dotyczy wystąpowania samego wyrażenia w jednym dokumencie,

Miary ważności

- *frequency weight* - dotyczy wystąpowania samego wyrażenia w jednym dokumencie,
- *term weight* - dotyczy liczby wystąpień danego wyrażenia w całej kolekcji dokumentów.

Miary ważności - frequency weight

- binarna – w_{ij} = 1 w przypadku wystąpowania zwrotu, a w_{ij} = 0 przypadku jego braku
- logarytmiczna - $w_{ij} = \log2(a_{ij} + 1)$ (logarytm przy podstawie 2 z liczby określającej częstość występowania słowa (a_{ij}))
  - pomniejsza wagę słów, które często się powtarzają;
- liczebnościowa - liczba wystąpowania słów bez modyfikacji: $w_{ij} = a_{ij}$.

Miary istotności - frequency weight

- transformacja *Okapi BM25*

$$w_{ij} = \frac{(k+1)a_{ij}}{a_{ij}+k},$$

Miary ważności - term weight

- *Normal* – waga ta jest proporcjonalna to liczby wystąpienia danego słowa w dokumencie;
- *None* – każdemu zwrotowi przypisuje się wagę 1;
- *Entropy* – przypisuje najwyższą wagę słowom, które wystąpiły najrzadziej w danym dokumencie;

Miary ważności - term weight

- *Chi-Squared* – wykorzystuje wartość testu Chi-kwadrat;
- *Mutual Information* – pokazuje jak rozkład dokumentów z wyrażeniem *i*, znajduje się blisko rozkładu dokumentów w całym zbiorze;
- *Information Gain* – określa oczekiwaną redukcję *Entropy* w przypadku podzieleniu zbioru dokumentów według tego wyrażenia *i*.

Miary ważności - term weight

- **IDF** (*Inverse Document Frequency*) waga jest odwrotnością liczby dokumentów, w których pojawił się dany zwrot;

$$IDF(term) = \log[\frac{m}{n}]$$

gdzie:

m – liczba dokumentów w korpusie

n – liczba dokumentów zawierających wskazany term

Miary ważności - TF-IDF

- **TF-IDF** (*term frequency-inverse document frequency*) polega na ustalaniu względnej częstotliwości słów w danym, lokalnym dokumencie i porównaniu z odwróconą częstotliwością słowa w całej kolekcji dokumentów.

$$TFIDF = TF * IDF$$

Miary ważności - TF-IDF

Strong tea is bad for a bad mood

Strong coffee is good for a bad mood

	dok_1	dok_2
strong	1	1
tea	1	0
is	1	1
bad	2	1
for	1	1
mood	1	1
coffee	0	1
good	0	1

Miary ważności - TF-IDF

Strong tea is bad for a bad mood

Strong coffee is good for a bad mood

	dok_1	dok_2	TF_1
strong	1	1	1/7 = 0,14
tea	1	0	1/7 = 0,14
is	1	1	1/7 = 0,14
bad	2	1	2/7 = 0,28
for	1	1	1/7 = 0,14
mood	1	1	1/7 = 0,14
coffee	0	1	0/7 = 0
good	0	1	0/7 = 0

Miary ważności - TF-IDF

Strong tea is bad for a bad mood

Strong coffee is good for a bad mood

	dok_1	dok_2	TF_1	IDF_1
strong	1	1	1/7 = 0,14	ln(2/2) = 0
tea	1	0	1/7 = 0,14	ln(2/1) = 0,7
is	1	1	1/7 = 0,14	ln(2/2) = 0
bad	2	1	2/7 = 0,28	ln(2/2) = 0
for	1	1	1/7 = 0,14	ln(2/2) = 0
mood	1	1	1/7 = 0,14	ln(2/2) = 0
coffee	0	1	0/7 = 0	ln(2/1) = 0,7
good	0	1	0/7 = 0	ln(2/1) = 0,7

Miary ważności - TF-IDF

Strong tea is bad for a bad mood

Strong coffee is good for a bad mood

	dok_1	dok_2	TF_1	IDF_1	$TFIDF_1$
strong	1	1	1/7 = 0,14	ln(2/2) = 0	0
tea	1	0	1/7 = 0,14	ln(2/1) = 0,7	0,098
is	1	1	1/7 = 0,14	ln(2/2) = 0	0
bad	2	1	2/7 = 0,28	ln(2/2) = 0	0
for	1	1	1/7 = 0,14	ln(2/2) = 0	0
mood	1	1	1/7 = 0,14	ln(2/2) = 0	0
coffee	0	1	0/7 = 0	ln(2/1) = 0,7	0
good	0	1	0/7 = 0	ln(2/1) = 0,7	0

Miary ważności - TF-IDF

	dok1	dok2	$TFIDF_1$	$TFIDF_2$
strong	1	1	0	0
tea	1	0	1/7*log(2)	0
is	1	1	0	0
bad	2	1	0	0
for	1	1	0	0
mood	1	1	0	0
coffee	0	1	0	1/7*log(2)
good	0	1	0	1/7*log(2)

TEXT MINING

Tekst mining

odkrywanie nieznanych wzorców, zależności z numerycznej postaci tekstu

Zadania Text Mining

- Klasyfikacja (kategoryzacja tekstu)
- Grupowanie (naturalne grupy tekstu)
- Analiza wydźwięku
- Analiza wątków

PYTANIA